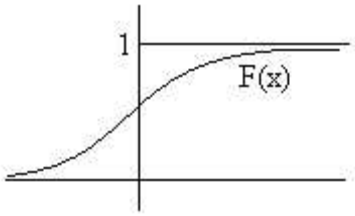
18.05 Lecture 35
May 9, 2005

Kolmogorov-Smirnov (KS) goodness-of-fit test
Chi-square test is used with discrete distributions.
If continuous - split into intervals, treat as discrete.
This makes the hypothesis weaker, however, as the distribution isn't characterized fully.
The KS test uses the entire distribution, and is therefore more consistent.

Hypothesis Test:
$H_1 : \mathbb{P} = \mathbb{P}_0$
$H_2 : \mathbb{P} \neq \mathbb{P}_0$
$\mathbb{P}_0$ - continuous
In this test, the c.d.f. is used.

Reminder: c.d.f. $F(x) = \mathbb{P}(X \leq x)$, goes from 0 to 1.

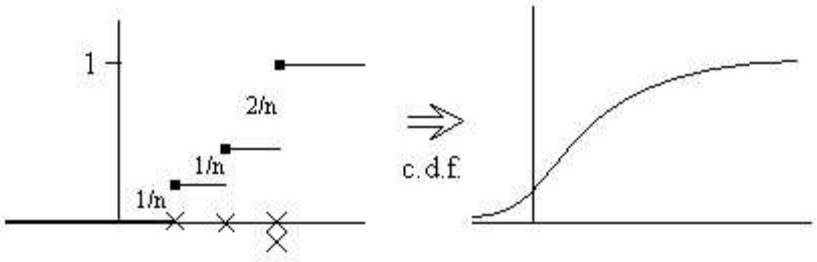


The c.d.f. describes the entire function.
Approximate the c.d.f. from the data →

Empirical Distribution Function:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X \leq x) = \frac{\#(points \leq x)}{n}$$

$$\text{by LLN, } F_n(x) \to \mathbb{E}I(X_1 \leq x) = \mathbb{P}(X_1 \leq x) = F(x)$$



From the data, the composed c.d.f. jumps by 1/n at each point. It converges to the c.d.f. at large n.
Find the largest difference (supremum) between the disjoint c.d.f. and the actual.

$$\sup_x |F_n(x) - F(x)| \overrightarrow{n \to \infty} 0$$

For a fixed x:

$$\sqrt{n}(F_n(x) - F(x)) = \frac{\sum(I(X_i \leq x) - \mathbb{E}I(X_1 \leq x))}{\sqrt{n}}$$

By the central limit theorem:

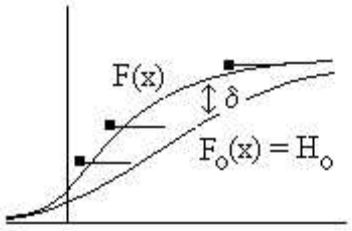
$$\approx N\Big(0, \text{Var}(I(X_i \leq x)) = p(1-p) = F(x)(1-F(x)\Big)$$

You can tell exactly how close the values should be!

$$D_n = \sqrt{n} \sup_x |F_n(x) - F(x)|$$

a) Under H_1, D_n has some proper known distribution.
b) Under $H_2, D_n \to +\infty$

If $F(x)$ implies a certain c.d.f. which is δ away from that predicted by $H_0 \to$



$F_n(x) \to F(x), |F_n(x) - F_0(x)| > \delta/2$
$\sqrt{n}|F_n(x) - F_0(x)| > \frac{\sqrt{n}\delta}{2} \to +\infty$

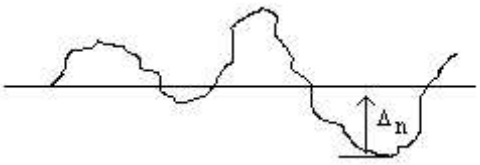
The distribution of D_n does not depend on F(x), this allows to construct the KS test.
$D_n = \sqrt{n} \sup_x |F_n(x) - F(x)| = \sqrt{n} \sup_y |F_n(F^{-1}(y)) - y|$
$y = F(x), x = F^{-1}(y), y \in [0, 1]$

$$F_n(F^{-1}(y)) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq F^{-1}(y)) = \frac{1}{n}\sum_{i=1}^{n} I(F(X_i) \leq y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y)$$

Y values generated independently of F.
$\mathbb{P}(Y_i \leq y) = \mathbb{P}(F(X_i) \leq y) = P(X_i \leq F^{-1}(y)) = F(F^{-1}(y)) = y$
$X_i \sim F(x)$
$F(X_i) \sim$ uniform on [0, 1], independent of Y.

D_n is tabulated for different values of n, since not dependent on the distribution.
(find table on pg. 570)
For large n, converges to another distribution, whose table you can alternatively use.
$\mathbb{P}(D_n \leq t) \to H(t) = 1 - 2\sum_{i=1}(-1)^{i-1}e^{-2i^2t^2}$

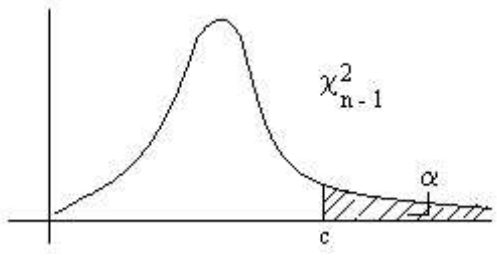
The function represents Brownian Motion of a particle suspended in liquid.



Distribution - distance the particle travels from the starting point.
The maximum distance is the distribution of D_n

H(t) = distribution of the largest deviation of particle in liquid (Brownian Motion)

Decision Rule:
$\delta = \{H_1 : D_n \leq c; H_2 : D_n > c\}$
Choose c such that the area to the right is equal to α



Example:
Set of data points as follows →
n = 10,
0.58, 0.42, 0.52, 0.33, 0.43, 0.23, 0.58, 0.76, 0.53, 0.64
$H_1 : \mathbb{P}$ uniform on [0, 1]
Step 1: Arrange in increasing order.
0.23, 0.33, 0.42, 0.43, 0.52, 0.53, 0.58, 0.64, 0.76
Step 2: Find the largest difference.
Compare the c.d.f. with data.

Note: largest difference will occur before or after the jump, so only consider end points.

x:	0.23	0.33	0.42	...
F(x):	0.23	0.33	0.42	...
$F_n(x)$ before:	0	0.1	0.2	...
$F_n(x)$ after:	0.1	0.2	0.3	...

Calculate the differences: $|F_n(x) - F(x)|$

$F_n(x)$ before and F(x):	0.23	0.23	0.22	...
$F_n(x)$ after and F(x):	0.13	0.13	0.12	...

The largest difference occurs near the end: $|0.9 - 0.64| = 0.26$
$D_n = \sqrt{10}(0.26) = 0.82$

Decision Rule:
$\delta = \{H_1 : 0.82 \leq c; H_2 : 0.82 > c\}$
c for $\alpha = 0.05$ is 1.35.
Conclusion - accept H_1.

** End of Lecture 35